Advances in Applied Mathematics 应用数学进展, 2023, 12(11), 4708-4716
Published Online November 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1211464

Hans汉斯

基于LightGBM模型的产品订单需求量预测

赵玉骁，陈思予，叶凯圳，施锋伟，金秀玲*

闽江学院数学与数据科学学院，福建 福州

收稿日期：2023年10月21日；录用日期：2023年11月14日；发布日期：2023年11月21日

摘　要

产品订单需求量预测是管理企业供应链的关键环节。准确预测客户对产品的需求量是很有必要的。为了解决不同产品的需求量问题，本文充分利用厂商数据，采用基于LightGBM的集成算法建立产品订单量预测模型。用网格探索进行参数调优，用3折目标编码，最终测试集上的MAPE为0.3541%；拟合效果良好且泛化能力强。最后用LightGBM模型预测后三个月的各个地区、各个品类的月度订单需求量。有助于企业资源有效配置，提高企业的收益效率，具有较大的现实意义和参考价值。

关键词

订单需求预测，机器学习，LightGBM

Product Order Demand Forecast Based on LightGBM Model

Yuxiao Zhao, Siyu Chen, Kaizhen Ye, Fengwei Shi, Xiuling Jin*

Colledge of Mathematics and Data Science, Minjiang University, Fuzhou Fujian

Received: Oct. 21st, 2023; accepted: Nov. 14th, 2023; published: Nov. 21st, 2023

Abstract

Product order demand forecasting is a key step in managing enterprise supply chain. It is necessary to accurately predict the customer's demand for the product. In order to address the demand forecasting challenge for different products, this article leverages vendor data extensively and employs an ensemble algorithm based on LightGBM to build a predictive model for product order

*通讯作者。

文章引用: 赵玉骁, 陈思予, 叶凯圳, 施锋伟, 金秀玲. 基于 LightGBM 模型的产品订单需求量预测[J]. 应用数学进展, 2023, 12(11): 4708-4716. DOI: 10.12677/aam.2023.1211464

quantities. Mesh exploration was used for parameter tuning and 3-fold target coding. The MAPE on the final test set was 0.3541%. The fitting effect is good and the generalization ability is strong. Finally, LightGBM model is used to forecast the monthly order demand of each region and category in the next three months. It is helpful to the effective allocation of enterprise resources and improve the profit efficiency of enterprises, and has great practical significance and reference value.

Keywords

Order Demand Forecasting, Machine Learning, LightGBM

1. 引言

1.1. 研究意义

随着经济全球化的高速发展，市场竞争愈发激烈。多种外部因素的不确定性使企业供应链面临着巨大的挑战。制造企业通常会生产一系列的产品，每个系列下又包含多种型号。这些型号产品之间相互制约影响，如何合理预测不同型号产品的需求量成为企业管理者需要解决的问题之一。目前，研究同一系列下多种型号产品之间的相互制约影响已经成为了热点问题，但是如何准确地预测不同型号产品的需求量仍然存在一定难度。此外，订单数据具有时序性，如何利用历史订单数据中的信息，建立能够准确预测需求量的模型，也是研究人员需要解决的问题。产品订单需求预测可以帮助企业管理供应链，准确预测产品订单需求能够支配生产、减少库存、降低成本、帮助资源配置，有利于公司制定产品销售及运营计划，从而制定合理的采购和生产计划，缩短订单交货时间，提高交易效率，增加客户满意度。

1.2. 文献综述

目前，国内外针对预测订单需求的问题已有较多的研究，主要分为传统统计学方法和现代机器学习方法两类。阚毅[1]运用多元回归分析方法，针对迅达公司产品订单需求进行预测。郭瑞[2]通过将定量预测的四种方法：简单平均法、加权移动平均法、一次指数平滑法、一元线性回归法进行比较，得到最优方法加权移动平均法来对 MTO 企业的订单进行预测，并进行订单排序，有效解决 MTO 企业的生产计划缺乏预测的问题。张崇娇、沈小林[3]等人采用果蝇算法和和灰色理论相结合，构建优化灰色神经网络的冰箱订单需求方法，提高订单需求的预测精度。曲艺[4]采用 BP 算法，针对 A 公司笔记本电脑需求预测的问题，基于 A 公司过去一年出货数据对 A 公司笔记本电脑的订单进行预测，并提出相应的库存管理意见。孙琳[5]构建了订单装配(ATO)生产模式下的产品订单的 NRA 时间序列预测模型，并于指数平滑法、移动平均法、灰色预测进行对比，得出 NRA 神经网络对于非线性、非稳定订单量时间序列来说预测效果更好。国外的学者 Kamala Aliyeva [6]通过添加 Z 信息量，建立 Z-回归模型的方式实现制造业需求量的预测。A Jayant，A Agarwal [7]等人使用了支持向量机的机器学习方法建立自回归模型预测摩托车订单量。

综上所述现有文献企业数据维度高、基数大，使用传统统计学的方法进行预测的效率较低；而使用深度学习的算法生成的模型又通常无法给出合理的解释。LightGBM 作为树模型，其本身具有较强的可解释性，并且使用该算法在处理样本量大、维度高的数据时，可以准确并快速地得到结果，解决了传统统

计学模型在处理企业海量数据时效率低、耗时长的问题。本文对产品的不同因素进行深入分析，通过建立基于 LightGBM 的模型来对订单需求量进行预测。有利于公司做出相应策略的结果，为优化供应链管理提供保障依据。

2. 相关理论

2.1. 需求预测相关理论

需求预测是指通过调查研究，充分利用已有数据，结合相关影响因素，寻找合适、科学的方法，建立恰当的数学模型，最后对未来需求发展趋势做出准确的判断[6]。需求预测对企业原材料采购、生产计划安排、库存管理、销售目标等方面都有重大影响，通过正确的需求预测，能够更好地制定企业未来的发展战略，从而使企业得到更有利的发展。

2.2. LightGBM 相关理论

虽然传统的 Boosting 算法(如 GBDT、XGBoost 等)已经有了相当好的效率，但传统的 Boosting 算法需要对每一个特征的每一个分裂点都要遍历所有的样本点计算信息增益从而选择最好的切分点，因此其计算复杂度将会受到特征数量和数据量的双重影响。这在现如今样本量和数据维度不断增大的环境下，其遍历次数多、限制数据大小和耗时长的弊端越发显现出来，这在面对工业级的海量数据时，使用普通的 Boosting 模型普遍是不能满足其需求的。

而 LightGBM (Light Gradient Boosting Machine) [6]作为一个实现 GBDT 算法的框架，可以有效地解决这种在大样本高纬度环境下的耗时问题。LightGBM 主要集成了两种算法单边梯度采样(GOSS)和互斥特征捆绑(EFB)对模型进行优化，同时使用了直方图算法做差加速。

(1) GOSS 算法

GOSS 是一种能够在减少数据量和模型精度上保持平衡的算法，其主要思想就是使梯度大的样本点在信息增益的计算上起主要的作用，即这些梯度大的样本点会贡献出更多的信息增益。为了保持信息增益评估的精度，对样本采样时应保留梯度大的样本点，并对梯度小的样本点按比例进行随机采样。

(2) EFB 算法

LightGBM 在对数据进行精简的同时，对特征也进行了降维处理，从而提高模型精度。高维的数据通常是稀疏的，而且这种稀疏特征空间中许多特征是互斥的，为了尽可能在不损失模型精度的同时减少特征维度，可以将互斥的特征绑定在一起成为同一个特征，这样做能够极大的加速 GBDT 算法的训练过程而且不会损失精度。

LightGBM 关于互斥特征的合并用到了直方图(Histogram)算法。其基本思想是先把连续的特征值离散化成 k 个整数，同时构造一个宽度为 k 的直方图。在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。

使用 GOSS 算法和 EFB 算法的梯度提升树称之为 LightGBM 集成算法。

2.3. 评价指标相关理论

本文使用了 RMSE 均方根误差，MAE 平均绝对误差和 MAPE 平均绝对百分比误差作为评价指标。RMSE 用于反映估计量与被估计量之间的差异程度；MAE 计算各测量值的绝对偏差的平均值，可以准确反映实际预测误差的大；MAPE 用于衡量时间序列值预测结果的准确性。

RMSE 均方根误差公式为：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2}$$

MAE 平均绝对误差公式为：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right|$$

MAPE 平均绝对百分比误差公式为：

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|$$

2.4. LightGBM 建模流程

LightGBM 模型的构建步骤如图 1 所示：将原始数据进行预处理后构建特征工程，创造新的特征因子用 LightGBM 模型训练。调整参数使模型达到最优，最后预测需求量，返回预测结果评价。

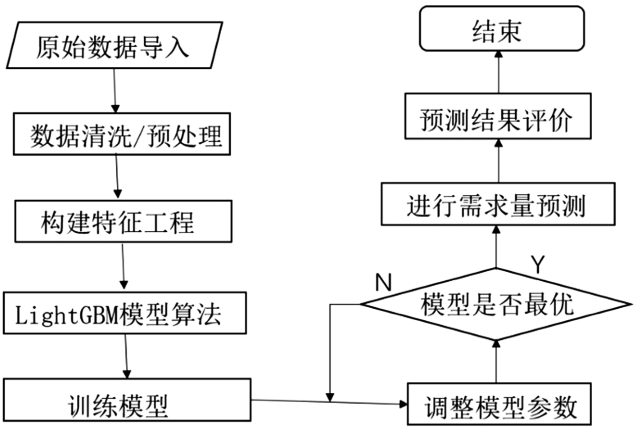

Figure 1. LightGBM modeling steps
图 1. LightGBM 建模步骤

3. LightGBM 实证分析预测订单需求量

3.1. 数据来源

本文选取了国内某大型制造企业 2015 年 9 月 1 日到 2018 年 12 月 20 日面向经销商的共 597694 条记录、8 个信息量，数据真实可靠。取 2015 年 9 月 1 日到 2018 年 8 月 31 日的出货数据和产品参数数据作为训练集训练模型，利用 2018 年 9 月 1 日到 2018 年 12 月 20 日的产品参数进行预测，预测不同品类产品的需求量。

对数据进行清洗，删除冗余数据和异常值并用均值插值法填补缺失值保证序列连续。由于产品需求量的时效性很强，在原有数据基础上，引入节假日、四季、时间段等时序特征。将经过数据预处理并新增特征因子的新数据集运用 LightGBM 模型计算对产品需求量影响较大的特征因子，选出特征重要性前 20 的特征因子参数组合到一起，进行标准化和编码处理，得到最终数据集。

3.2. 描述性数据分析

将通过以上处理得到的数据合并成新数据集，并对其进行描述性统计分析，得到结果如表 1 所示。

Table 1. Descriptive data analysis table
表 1. 描述性数据分析表

特征	平均值	方差	最小值	25%分位数	50%分位数	75%分位数	最大值
地区	2.904	2.265	1	2	3	5	5
大类	304.932	4.019	301	303	306	306	308
细类	406.5	9.659	401	404	407	408	412
销售方式	0.265	0.195	0	0	0	1	1
价格	10722.68	1077102.62	11	5982	8843	12914	980164
日	200.6	37056.739	1	107	218	296	366
需求量	91.46	11415.84	1	10	29	101	10547

考虑到节假日对产品需求量也会有一定影响，因而本文选取了国内的法定节假日进行分析，对比节假日和非节假日的订单需求量，并进行可视化分析如图 2 所示：

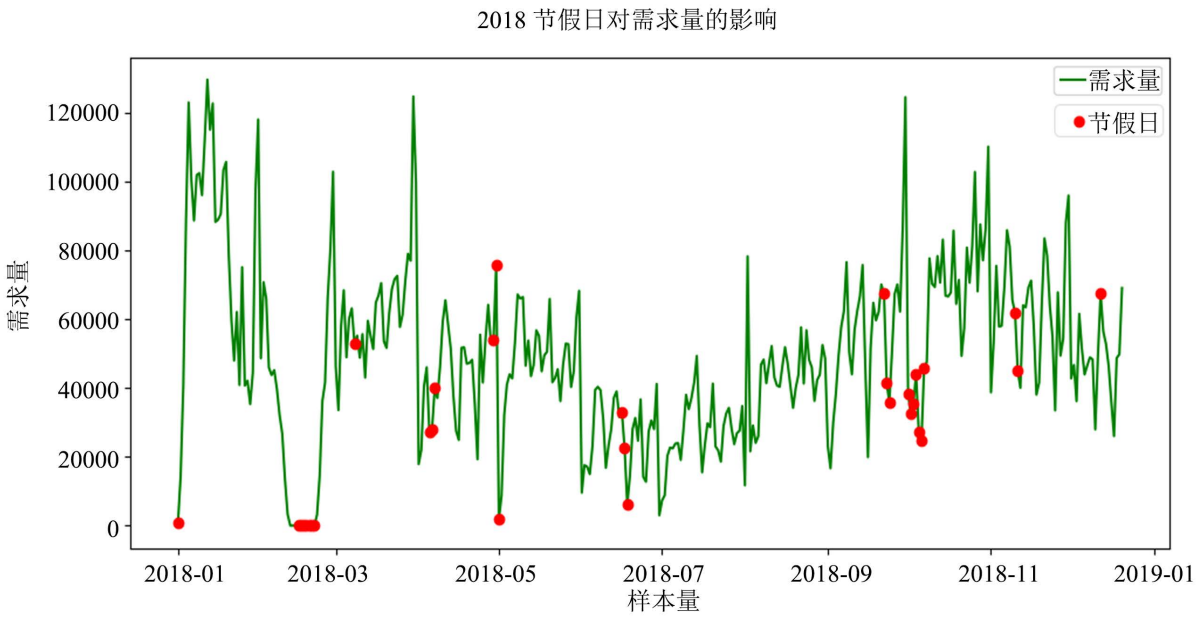

Figure 2. Line chart of holiday product demand in 2018
图 2. 2018 年节假日产品需求量折线图

图 2 选取了 2018 年的数据分析其节假日的和非节假日的产品订单需求量，非节假日的订单需求量远远高于节假日的订单需求量。这表明在节假日产品订单需求会降低，可能是由于法定节假日企业大多休假，导致订单出货数量减少。

不同的商品在不同季节会有不同的需求量，所以季节因素对产品订单销售需求量也会造成一定的影响，对时间序列提取出季节因素并绘制出图 3。

可以看到，冬天的订单需求量要高于其他季节，而且会随着温度的升高而降低。考虑到可能是由于冬天的促销和节假日较多，消费需求也会增高。

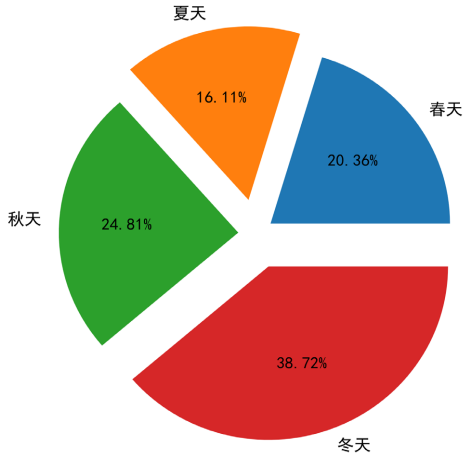

Figure 3. Demand proportion diagram by season
图 3. 各季节需求比例图

3.3. LightGBM 模型构建

以 2015 年 9 月 1 日到 2018 年 8 月 31 日的数据作为训练集，2018 年 9 月 1 日到 2018 年 12 月 20 日共 111 天的数据作为验证集，分别构建 XGBoost、CatBoost 和 LightBGM 模型，可视化预测值与真实值的对比图，检验模型预测效果。

3.1.1. LightGBM 模型参数调优

本文所选用的模型参数众多，对模型精度影响较大，参数选择不恰当出现欠拟合或者过拟合的问题，为了提高模型精度，同时保证模型的泛化能力，调参是模型建立中不可或缺的一部分，对于参数的选择，主要选择网格搜索配合经验微调调参，可以保证在指定参数范围内找出精度最高的参数。

基本调参过程如下：首先选取较高的学习率，数值在 0.1 附近，这么做可以加快收敛速度，对于其他参数调整很有必要。其次是对决策树基本参数的调整，可以有效提高模型精度，最后调整正则化参数，可以防止模型过拟合。因此，在选定较高的学习率后，第一步确定数最大深度和决策树数目，第二步确定最小叶子节点样本权重和特征值离散化分段数，这一步是为了防止过拟合。第四步确定样本采样率，这一步可以用来加速训练并且处理过拟合数据。第五步确定 L1 正则化参数和分裂最小增益阈值。其他参数根据网格搜索遍历参数选择最优参数，最后降低学习率，构建最优预测模型。

LightGBM 模型的最终调参结果如表 2 所示。

Table 2. LightGBM parameter list
表 2. LightGBM 调参表

参数名称	参数范围	参数选择	评价指标
决策树最大深度	[3,15]	8	AUC: 91.42% MAPE: 0.3541% Time: 1 m 21 s
决策树数目	[1000,30000]	20000	
最小叶子节点样本权重	[1,8]	5	
特征值离散化分段数	[5,255]	205	
样本采样率	[0,2,2,0.1]	1	
L1 正则化参数	[0,1]	0.1	
学习率	[0.005,0.1]	0.05	
分裂最小增益阈值	[1,91]	76	

用最优参数建模对验证集 20 天的数据进行预测，AUC 有 91.42%，很接近 1.0。与真实值对比，最大的 MAPE 是 7.118%，最小 MAPE 是 0.0012%，平均 MAPE 为 0.3541%，从图 4 上可以看出误差整体波动幅度较小，具有较好的拟合效果。

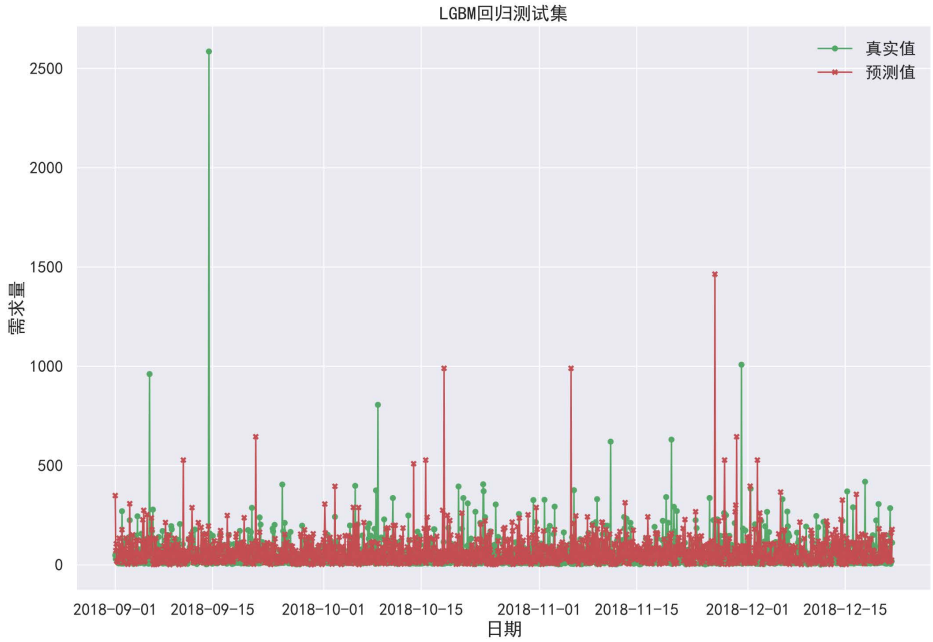

Figure 4. LightGBM fitting diagram
图 4. LightGBM 拟合图

3.3.2. 各模型验证集效果对比

XGBoost 模型和 CatBoost 模型与 LightGBM 模型和真实数据的对比图和评价指标表如表 3 所示。

Table 3. Comparison table of model prediction effect
表 3. 模型预测效果对比表

模型	验证集			运行时间
	MAE	RMSE	MAPE (%)	
XGBoost	3052.32	3433.21	0.843	1 m 32 s
CatBoost	1895.19	2452.56	0.7232	1 m 36
LightGBM	1094.87	1946.77	0.3541	1 m 21 s

通过上述不同模型拟合结果和评价指标对比可知，XGBoost 模型的预测精确度最低，预测值与真实值的拟合度最差，CatBoost 模型次之，LightGBM 模型预测效果最好。相比 XGBoost 和 CatBoost，LighrGBM 的 MAE 分别减少 1957.45 和 800.32，RMSE 分别减少 1486.44 和 505.79，MAPE 分别减少 0.4889%和 0.3691%，在训练集上，LightGBM 拟合效果最好，CatBoost 其次，最差的是 XGBoost，各项指标几乎是 LightGBM 的 2.5 倍，从运行时间看，CatBoost > XGBoost > LightGBM，三种模型都是基于集成的决策树模型，因此运行时间的差距并不大，影响运行时间的主要因素是学习率和数据特征分割数，模型本身在处理大规模数据集时具有优势，从中得出，LightGBM 在预测方面表现出明显的优势，可以更好地进行本文数据的预测。

3.3.3. 预测 2019 年 1 月至 3 月的需求量

因此，LightGBM 适用于产品订单的需求量预测，我们基于上述参数范围内对 2019 年 1 月到 3 月各销售地区一千七百多种的品类分别进行月度需求量的预测，从而可以预测出每个月的产品订单需求量。企业可以根据预测的订单需求量预生产或安排生产计划。举产品品类 20002 来说明预测的结果，如表 4。

Table 4. Product 20002 forecast overview
表 4. 产品 20002 预测概况

销售地区	产品品类	大类编码	细类编码	一月	二月	三月
101	20002	303	406	662	1208	1240
102	20002	303	406	1265	941	203
103	20002	303	406	844	1034	1423
104	20002	303	406	0	0	0
105	20002	303	406	746	1660	2028
总和				3517	4843	4894

表 4 为产品品类 20002 在 LightGBM 模型下预测 2019 年 1 月至 3 月的月度订单需求量，可以看出在不同销售地区，同样品类产品有不同的订单需求量。利用 LightGBM 模型预测可以帮助厂商提早指定生产计划，优化产业链。

4. 总结

本文通过结合企业面向经销商的各类产品需求量的历史数据与机器学习算法进行模型预测，主要目的是构建最优拟合模型预测经销商对各类产品需求量的短期预测。最终得到如下结论：

(1) 本文将时间特征数据和产品数据按照销售地区和产品编码分组合并，将合并后的特征因子放入机器学习算法中预测未来 20 天的产品需求量，通过对模型训练集的 MAPE 的计算，模型得到了较好的预测效果，可见按本文的方法预测具有一定和合理性和正确性。

(2) 在采用 XGBoost、CatBoost、LightGBM 三种机器学习算法最优调参后，LightGBM 的 MAPE 值分别减少 0.4889%和 0.3691%，效果好于其他算法，预测数据的精度更高，预测值与真实值的拟合度也更高，在预测能力上表现出明显的优势。

本文能利用历史数据对商品需求量进行预测，可以帮助产家有效了解供求关系，制定生产计划，优化供应链。基于历史数据对未来短期数据预测思路清晰，算法具有可迁移性，可以作为其他场合的数据预测的参考。

基金项目

项目来源：福建省科技厅；
项目名称：基于 ATOT 技术的智能养老系统设计与开发；
项目编号：2023350104000282。

参考文献

[1] 阚毅. 迅达公司产品订单预测模型研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2015.
[2] 郭瑞. MTO 企业订单需求确定与排序研究[D]: [硕士学位论文]. 北京: 中国矿业大学, 2022.
[3] 张崇娇, 沈小林, 等. 基于果蝇算法优化灰色神经网络的冰箱订单需求预测研究[J]. 数学的实践与认识, 2017,

47(20): 15-19.

[4] 曲艺. A 公司的需求预测与库存管理改善研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2019.

[5] 孙琳. 基于 ATO 的供应链采购优化应用研究[D]: [硕士学位论文]. 南京: 东南大学, 2020.

[6] Aliyeva, K. (2017) Demand Forecasting for Manufacturing under Z-Informatio. *Procedia Computer Science*, **120**, 509-514. https://doi.org/10.1016/j.procs.2017.11.272

[7] Jayant, A. and Agarwal, A. (2020) Application of Machine Learning Technique for Demand Forecasting: A Case Study of Manufacturing Industry. In: Pandey, P.M., Kumar, P. and Sharma, V., Eds., *Advances in Production and Industrial Engineering*, *Lecture Notes in Mechanical Engineering*, Vol. 1236, Springer, Singapore, 403-421. https://doi.org/10.1007/978-981-15-5519-0_31